ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-6327-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICASSP49357.2023.10094998

INV-SENET: INVARIANT SELF EXPRESSION NETWORK FOR CLUSTERING UNDER BIASED DATA

Ashutosh Singh[i,*], Ashish Singh[j,*], Aria Masoomi[i], Tales Imbiriba[i], Erik Learned-Miller[j], Deniz Erdoğmuş[i]

[i] Dept. of Electrical & Computer Engineering, Northeastern University, Boston, MA, USA
[j] College of Information & Computer Sciences, University of Massachusetts Amherst, MA, USA

ABSTRACT

Subspace clustering algorithms are used for understanding the cluster structure that explains the patterns prevalent in the dataset well. These methods are extensively used for data-exploration tasks in various areas of Natural Sciences. However, most of these methods fail to handle confounding attributes in the dataset. For datasets where a data sample represent multiple attributes, naively applying any clustering approach can result in undesired output. To this end, we propose a novel framework for jointly removing confounding attributes while learning to cluster data points in individual subspaces. Assuming we have label information about these confounding attributes, we regularize the clustering method by adversarially learning to minimize the mutual information between the data representation and the confounding attribute labels. Our experimental result on synthetic and real-world datasets demonstrate the effectiveness of our approach.

Index Terms— clustering, bias mitigation, subspace

1. INTRODUCTION

Most real-world datasets carry information arising from several attributes. Given the task of estimating some unknown value, such as the category of input, some of these attributes have no information about the task, i.e., they are statistically independent of the desired value. Ideally, given large enough samples taken iid from the data distribution, these attributes should be uncorrelated to the features of the dataset that are important to fulfill the task. However, some attributes can still be highly correlated with the informative features in many real-world datasets due to non-iid sampling or data collection procedures [1]. When training data-driven models on such datasets, these attributes can negatively affect the inference results. We refer to such attributes as confounding attributes or biases.

While many past works have proposed to learn models which are robust to the presence of confounding attributes (bias mitigation), they generally address supervised learning tasks [2][3][4], where task-relevant label information is available during model training time. However, bias mitigation strategies for the unsupervised setting, i.e., when task-relevant information is absent, have largely been understudied. In this work, we address the problem of unsupervised learning when data can contain confounding attributes.

Among the data-driven approaches for unsupervised learning, subspace clustering shows great promise. In subspace clustering, the assumption is that the high dimensional data lie on a union of low dimensional subspaces. The objective here is to find separate subspaces for separate clusters of data points. Among the family of subspace clustering algorithms, Self-expression based algorithms form the state of the art[5]. Self-expression imposes the constraint that every data point in the dataset can be explained through a linear combination of all the other data points in the dataset. Let c_{ij} represents the coefficient of the j^{th} datapoint w.r.t. its contribution in reconstructing the i^{th} datapoint. One of the reasons why self-expression-based methods are popular is because of the subspace preserving property of the coefficient matrix achieved under certain regularisation function [6][7][8][9][10]. This means that c_{ij} is only non-zero when the i^{th} and j^{th} data points are in the same subspace. Most recent advances in subspace clustering literature have focused on scalability and out-of-sample clustering using neural networks. But even these algorithms fail under the presence of confounding attributes, as we later show.

More often, we have labels pertaining to these confounding attributes. Under the assumption that the labels of the biases are known during the training, we propose an information-theoretic inspired method of jointly learning the cluster membership while ignoring the confounding attributes in the data. We evaluate our proposed method over synthetic and naturally occurring image datasets and show superior performance than the current state of the art.

The remaining paper is organized as follows. Section 2 presents the problem statement, section 3 presents the proposed solution, section 4 presents the experiments and results, and section 5 presents the conclusion.

2. PROBLEM STATEMENT

Most clustering algorithms must be trained on the whole dataset to find clusters. As the size of the dataset increases, they become hard to scale. Scalable Subspace clustering algorithms show great promise in solving this problem [11][12][13]. The method proposed in [5], based on self-expression, scales well for large datasets and shows great performance on out-of-sample data points. But under the presence of confounding attributes, these methods fail on out-of-sample data points resulting in learning wrong self-expression coefficients. This paper focuses on this problem and looks toward improving the existing method. Below we formulate the problem of clustering under the presence bias information.

2.1. Formulation

Let $x \in \mathbb{R}^d$ be the input data and $X = [x_1, \ldots, x_n] \in \mathbb{R}^{d \times n}$ represent the input data matrix where the x_j represents different data points while the rows represent the dimensions. We make the assumption that the data, i.e., the columns in X lie on the union of k low-dimensional subspaces $x_j \in \cup_{i=1}^{k} X_i$. Let $b_j = [b_1, \ldots, b_m] \in B$ for m biases, $b_i \in \{0, 1\}$, we define $B = [b_1, b_2, \ldots, b_n]$ as

This work was supported by NSF (1947972).

[*] Indicates shared first authorship. Correspondence should be addressed to Ashutosh Singh (singh.ashu@northeastern.edu).

the matrix containing bias labels corresponding to X for n samples. Based on the self-expressive models [6], we can define each data point in X as a linear combination of all the other data points as [6],

$$x_j = \sum_{i \neq j} c_{ij} x_i \tag{1}$$

Here $c_{ij} \in \mathbb{R}$ represents the elements of the matrix C, self-expressive coefficient matrix, satisfying the following,

$$X = CX, \quad \text{diag}(C) = 0 \tag{2}$$

Hence learning of the self-expressive coefficient matrix could be formulated as the following optimization problem,

$$\min_C L(X, CX) + R(C), \tag{3}$$

where L and R represent cost and regularization terms, respectively. Here $R(\cdot)$ enforces subspace preserving property in the learned C [14]. Therefore the C matrix, subspace preserving, can be further used to get the affinity matrix $A = |C| + |C^T|$. We can obtain the clustering labels s over X using the affinity matrix in the spectral clustering algorithm [15].

Now, let s be the random variable representing the set of true (desired) cluster labels, and $I(x, x')$ be the mutual information between random variables x and x'. Suppose the random variable $\hat{s}$ represents the cluster labels based on the C matrix we get from solving (3). Then due to the presence of confounding attributes in the data, we can very possibly observe,

$$I(\hat{s}, b) \not\approx I(s, b) \tag{4}$$

The effect of confounding attributes on classification task is already well discussed and studied in literature [2]. Therefore in tasks where supervision is not present, the strong presence of confounding attributes in the data would lead to clustering outcome getting biased too as formulated in (4). This would likely result in $\hat{s} \neq s$. In this contribution, we aim at solving the problem in (3) while enforcing bias mitigation as expressed in Eq. (4). To solve this problem we propose an information theoretic based approach in the next section.

3. BIAS MITIGATION FOR SELF-EXPRESSIVE NETWORKS

In this section, we discuss the elements of our proposed solution and finally discuss the architecture and optimization for Inv-SENet.

3.1. Self Expressive Network

In subspace clustering literature, there are many optimization frameworks already proposed that try to solve (3) to learn the self expression coefficient matrix C. Past works have tried to solve (3) by optimizing the following loss,

$$L_{SE} = \min_{c_{ij}, i \neq j} \|x_j - \sum_{i \neq j} c_{ij} x_i\|_2^2 + \sum_{i \neq j} r(c_{ij}) \tag{5}$$

where $r(\cdot)$ represents the regularisation term based on elastic net regularisation [14]. The role of $r(\cdot)$ is to provide a denser subspace preserving coefficient matrix. Let $f(\cdot): \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be the function, parameterised by Θ taking arguments x_i and x_j, and giving self-expression coefficients $c_{i,j}$ in matrix C, i.e. $f(x_i, x_j; \Theta) = c_{ij}$. Therefore we can rewrite (5) as the following optimisation,

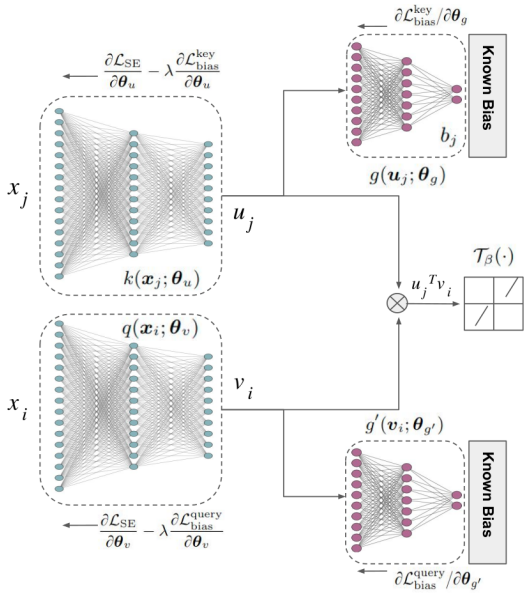


Fig. 1. Architecture for Inv-SENet. Here $k(x; \theta_u)$ and $q(x; \theta_v)$ be key and the query net respectively. $g(u_j; \theta_g)$ and $g'(v_i; \theta_{g'})$ are the bias classifiers.

$$L_{SE} = \min_{\Theta} \frac{\gamma}{2} \|x_j - \sum_{i \neq j} f(x_i, x_j; \Theta) x_i\|_2^2 + r(f(x_i, x_j; \Theta)) \tag{6}$$

In [5], the authors propose the SENet, which models $f(x_i, x_j; \Theta)$ as neural network expressed as

$$f(x_i, x_j; \Theta) = \alpha T_\beta(u_j^T v_i)$$
$$u_j = k(x_j; \theta_u) \in \mathbb{R}^p$$
$$v_i = q(x_i; \theta_v) \in \mathbb{R}^p$$

where,

$$T_\beta(\cdot) = \text{sign}(t) \max(0, |t| - \beta). \tag{7}$$

Here k and q represent the key and the query networks, while T_β is a soft thresholding operator. $\Theta = \{\theta_u, \theta_v, \beta\}$ represents the learnable parameters of the model. We highlight that SENet is not designed to cluster the data when bias is present in the data. We improve upon their model by adding a bias mitigation mechanism. The resulting methodology leads to a learning algorithm capable of performing clustering and mitigation jointly as described next.

3.2. Bias Mitigation

As discussed in Section 2.1, the presence of confounding attributes in the training dataset leads to bias-dependent solutions for the optimization problem and, therefore, leads to biased clustering. Below we present the proposed bias mitigation strategy. For this, we assume that the training dataset contains labels for the confounding attribute classes. Note, however, that these labels are not required for the test set. Therefore the goal here is to propose a way of learning Θ s.t. c_{ij} is invariant to the presence of confounding attributes. Referring to the proposed architecture in Fig. 1 we define $g(k(x; \theta_u); \theta_g): \mathbb{R}^p \to B$ and $g'(q(x; \theta_v); \theta_{g'}): \mathbb{R}^p \to B$ as the bias classification functions. Let u and v be the random variable representing the output of key and query networks, and random

variable b represent the confounding attribute. Therefore the optimisation framework in (6) can be modified as:

$$L = L_{SE} - \lambda I(u, b) + I(v, b). \quad (8)$$

Here $\lambda \in \mathbb{R}^+$ is a scalar hyperparameter controlling the bias mitigation component in the loss. Estimating $I(\cdot)$ requires the joint distributions $p(b, v)$ and $p(b, u)$. This makes the problem hard to solve and often cases intractable. Mutual information can also be written in terms of the conditional entropy such that

$$I(b, v) = H(b) - H(b|v), \qquad I(b, u) = H(b) - H(b|u).$$

Here, $H(b) = -\sum_{b_i \in B} P(b_i) \log P(b_i)$, represents the entropy. Note that $H(b)$ can be ignored when optimizing (8) since the bias b is independent of the model parameters Θ. To compute conditional entropies conditional densities such as $P(b|u)$ and $P(b|v)$ are required but can be difficult to obtain. In [2] the authors approximated the conditional distribution using a variational distribution Q. Following a similar reasoning we can define two distributions Q and Q^1 to approximate $P(b|u)$ and $P(b|v)$, respectively. Now, let θ_g and θ_{g^1} be the parameters of Q and Q^1. Thus, each parameter can be learned by minimizing the Kullback–Leibler (D_{KL}) divergence between the respective distributions. For θ_g we have:

$$\min_{\theta_g} D_{KL}(Q(g(k(x; \theta_u); \theta_g)|k(x; \theta_u)) \| P(b|k(x; \theta_u))) \quad (9)$$

while an equivalent problem can be solved for $v = q(x; \theta_v)$, omitted here for sake of space.

Now, we can define the bias classification optimization problem by combining a cross-entropy loss and the KL in (9) as a regularization term. The goal of this optimization is not only to make the bias classifier able to classify the bias class accurately but also to make the $k(x; \theta_u)$ invariant of the bias. To solve the second objective we can impose a regularisation on θ_u aiming at maximizing the negative of (9), hence making it hard for the $g(\cdot)$ to classify the bias. Therefore the optimization becomes a min max problem,

$$L^{key}_{bias} = \min_{\theta_u} \max_{\theta_g} E_{x \sim P_X p(\cdot)} [E_{b \sim \tilde{Q} p(\cdot)|k(x;\theta_u)} [\log Q(b|\tilde{k}(x; \theta_u))]] - \mu E_{x \sim P_X p(\cdot)} [L_c(b, g(k(x; \theta_u); \theta_g))] \quad (10)$$

where $L_c(\cdot)$ represents the cross entropy loss. The second term in (10) is the relaxation term for the KL divergence condition in (9). Similar discussion is presented in [2]. Equivalently, we can define the minimax problem L^{query}_{bias} for optimizing θ_v and θ_{g^1}, also omitted in this manuscript.

Let $\Theta_g = \{\theta_g, \theta_{g^1}\}$ represent the set of all parameters of the bias classifiers. Therefore, combining (10), for both key and query networks, and (5) we can rewrite (8) as:

$$\begin{aligned} L &= L_{SE} - \lambda(L^{key}_{bias} + L^{query}_{bias}) \\ &= \min_{\Theta} \max_{\Theta_g} \frac{\gamma}{2n} \sum_j \{x_j - \sum_{i \neq j} T_\beta(k(x_j; \theta_u)^T q(x_i; \theta_v)) q(x_i)\}_2^2 \\ &\quad + r(T_\beta(k(x_j; \theta_u)^T q(x_i; \theta_v)), q(x_i)) \\ &\quad - \lambda E_{x \sim P_X p(\cdot)} [E_{b \sim \tilde{Q} p(\cdot)|k(x;\theta_u)} [\log Q(\tilde{b}|k(x; \theta_u))]] \\ &\quad + \mu E_{x \sim P_X p(\cdot)} [L_c(b, g(k(x; \theta_u); \theta_g))] \\ &\quad - \lambda E_{x \sim P_X p(\cdot)} [E_{b \sim \tilde{Q}^1 p(\cdot)|q(x;\theta_v)} [\log Q^1(\tilde{b}|q(x; \theta_v))]] \\ &\quad + \mu E_{x \sim P_X p(\cdot)} [L_c(b, g^1(q(x; \theta_v); \theta_{g^1}))]. \end{aligned} \quad (11)$$

3.3. Training

In (11) we present the overall loss function for the architecture proposed in Fig. 1. In practice, we use two optimizers; one to solve the inner maximization and the second one to solve the outer minimization. First we compute the forward pass of the data through $k(x_j; \theta_u)$ and $q(x_i; \theta_v)$ and compute L_{SE}. At the same time, we also pass both u_j and v_i through their respective bias classifiers $g(\cdot)$ and $g^1(\cdot)$ to get the bias classification. We then compute the $BL^{key}_{bias}/B\theta_g$ and $BL^{query}_{bias}/B\theta_{g^1}$ and backpropagate through the bias classifiers $g(\cdot)$ and $g^1(\cdot)$ respectively. Using the gradient reversal technique [16] together with the gradient of L_{SE} we compute,

$$\nabla\theta_u = \frac{BL_{SE}}{B\theta_u} - \lambda \frac{BL^{key}_{bias}}{B\theta_u}, \qquad \nabla\theta_v = \frac{BL_{SE}}{B\theta_v} - \lambda \frac{BL^{query}_{bias}}{B\theta_v}.$$

for key and query networks, respectively. At the beginning of the training, the bias classifiers converge quickly. But as the training goes on, the C matrix becomes invariant of the biases, and the bias classifier starts performing poorly. This happens because the $I(u, b)$ and $I(v, b)$ goes down i.e the key, and the query net becomes good at unlearning the biases. The proposed method has constant memory requirement similar to SENet and is comparable in runtime.

4. EXPERIMENT

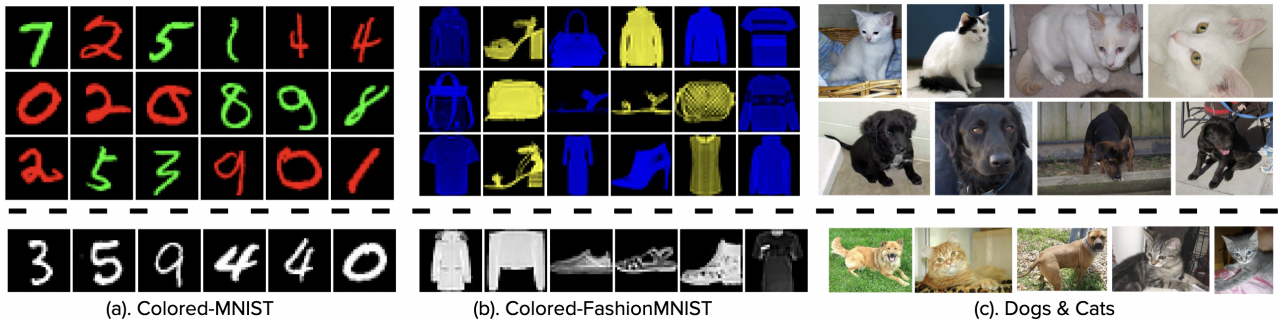
We conduct experiments on real-world benchmarks to evaluate our model performance. Specifically, we utilize three image datasets that contain confounding attributes to evaluate our method.

4.1. Setup

Network Architecture: We formulate query, key, and bias networks as an MLP. For query and key network, we create a three-layer MLP with ReLU and tanh(·) as the activation functions. The number of hidden units in each layer of these MLPs are {1024, 1024, 1024}, and the output dimension is 1024. For the Bias networks, we use a three-layer MLP of dimensions {1024, 512, 256} followed by the classification layer. We use the ReLU activation function and batch normalization layer between each fully-connected layer. We apply the softmax function to the output of the classification layer to compute class probability scores. To optimize our model, we use Adam [17] optimizer with a constant learning rate of $1e^{-3}$ for the query and key networks and $1e^{-4}$ for the bias networks.

Datasets: To evaluate our method, we consider two scenarios where our data samples contain confounding attributes. In the first setup, we intentionally add bias information to existing benchmarks (MNIST [18], and FashionMNIST [19]). In the second case, we consider a natural setting (Dogs and Cats dataset) where the bias information is inherently present in the data as a confounding attribute. Figure2 shows some of the samples from our constructed datasets.

Setting 1:We evaluate the proposed method on MNIST and FashionMNIST datasets as our representative large-scale scale benchmarks to evaluate our method. Both benchmarks consist of grayscale images. We color each image red or green for MNIST and blue and yellow for FashionMNIST to introduce confounding attribute. We color each image in a manner that strongly correlates with original class labels [20]. This allows us to measure the generalization properties of a given method effectively. If self-expressive coefficients are directly learned using the data without handling the confounding attributes, it should fail to generalize to out-of-distribution samples where these confounding attributes might be absent. We create our



(a). Colored-MNIST (b). Colored-FashionMNIST (c). Dogs & Cats

Fig. 2. Examples of Datasets with bias [Top-Rows: Training set; Bottom-Rows: Test set]. We modified the MNIST and FashionMNIST datasets by randomly adding color ([Red, Green] for MNIST and [Blue, Yellow] for FashionMNIST) to each dataset. We sub-sampled dark dog and bright cat images for the Dogs and Cats dataset.

training set in the following manner: We first assign a binary label y to the image based on the digit/class i.e. we set y " 0 for digits 0 ´ 4 (classes 0 ´ 4 for FashionMNIST) and y " 1 for digits 5 ´ 9 (classes 5 ´ 9 for FashionMNIST). We then flip the label with 25% probability. After that, we color the image as one of the colors according to its (possibly flipped) label. We flip the color with a probability e. For both benchmarks, we set e " 40%. For testing, we use the test sets of the original datasets without any confounding attribute. We use scattering convolution transform[21] to generate features from MNIST, and FashionMNIST similar to [5] for optimal comparison.

Setting 2: We select the Dogs and Cats dataset from [22] for evaluation on the benchmark where confounding attributes naturally exists. Due to the complex nature of the background and non-uniformity of the colour attribute across the images, bias mitigation task becomes challenging and non trivial. Hence benchmarking on this dataset helps us test the efficacy or the proposed method. We followed the similar setting from [2], wherein we subsample the original training set consisting of 25K images to create two subsets: a biased subset with dark-colored dog images and bright-colored cat images and a test set of images that does not contain any dark or bright colored dog or cat images. We obtained 6378 images as our train set and 8125 images as the test set. Thus in this setting, the bias classes are tdark, brightu. For feature extraction, we used [23] as it learns to represent features in a union of subspaces.

Metrics: For quantitative evaluation, we consider clustering accuracy (ACC), normalized mutual information (NMI) [24] and adjusted rand index (ARI) [25]. These metrics are commonly used in the literature to evaluate clustering methods.

Hyperparameter Selection:We selected hyperparameters μ and λ using random search and the 5-fold cross-validation (CV) method. We trained our proposed method for each CV iteration on 4 randomly sampled subsets and evaluated the clustering performance on the held-out subset using the NMI metric. The final values that we selected for each dataset are as follows.(i) MNIST: μ " 0.1, λ " 0.01; (ii) Fashion-MNIST: μ " 0.1, λ " 0.02; (iii) Dogs and Cats: μ " 0.3, λ " 0.05.

4.2. Results

We show our experimental results in Table 1. We evaluate the vanilla SENet and Inv-SENet models for out-of-distribution clustering. Here N_i represents the number of training samples. In this experiment, we want to test the generalizability of the learned self-expression coefficient matrices. To this end, we first train the model on the biased version of each dataset and then test it on the unbiased version. From the results in table 1, we can observe that having confounding attributes in a dataset can negatively affect clustering performance if the clustering is not invariant to such attributes. This is evident from the results of the standard SENet model, which fails to accurately cluster MNIST and FashionMNIST data based on ground-truth categories. The model performance is only marginally better than random chance (10% clustering accuracy for MNIST and FashionMNIST; 50% clustering accuracy for Dogs and Cats). In contrast, we see considerable improvement of our proposed method over the standard model. In particular, we see a significant increase in all three metrics for MNIST, while for FashionMNIST, we see moderate improvements. For the Dogs and Cats dataset, where the confounding attributes are naturally present, we again see improvement in clustering performance when incorporating invariance to the confounding attribute. This demonstrates that our method is able to effectively mitigate non-trivial confounding information and hence provide more accurate data clusters without any prior knowledge on the interaction between confounding attributes and useful features.

Table 1. Out-of-distribution clustering performance

Dataset	Method	N_i	ACC (%)	NMI. (%)	ARI (%)
MNIST	SENet	10000	41.79	33.34	21.05
		60000	36.36	25.14	16.14
	Inv-SENet	10000	68.44	60.55	45.23
		60000	78.04	67.53	58.56
Fashion-MNIST	SENet	10000	38.47	30.15	19.51
		60000	46.73	42.46	31.18
	Inv-SENet	10000	50.73	42.15	30.14
		60000	56.91	44.68	34.41
Dogs and Cats	SENet	6738	64.07	07.59	06.94
	Inv-SENet	6738	78.53	25.32	32.4

5. CONCLUSION

In this work, we formulated a bias mitigation strategy aimed at learning confounding attribute (bias) invariant self-expression coefficients. The proposed Inv-SENet is effective in learning bias-invariant subspace clustering for data under the presence of confounding attributes. Through experiments on synthetic and real world data we demonstrate the effectiveness of the proposed method. Future work will focus on extending this work to datasets with modalities having multiple confounds such as EEG[26], speaker recognition[27] and fMRI[28][29].

6. REFERENCES

[1] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database forstudying face recognition in unconstrained environments," in Workshop on faces in'Real-Life'Images: detection, alignment, and recognition, 2008.

[2] B. Kim, H. Kim, K. Kim, S. Kim, and J. Kim, "Learning not to learn: Training deep neural networks with biased data," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 9012–9020.

[3] M. Alvi, A. Zisserman, and C. Nellåker, "Turning a blind eye: Explicit removal of biases and variation from deep neural network embeddings," in Proceedings of the European Conference on Computer Vision (ECCV) Workshops, 2018, pp. 0–0.

[4] J. Attenberg, P. Ipeirotis, and F. Provost, "Beat the machine: Challenging humans to find a predictive model's "unknown unknowns"," Journal of Data and Information Quality (JDIQ), vol. 6, no. 1, pp. 1–17, 2015.

[5] S. Zhang, C. You, R. Vidal, and C.-G. Li, "Learning a self-expressive network for subspace clustering," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 12 393–12 403.

[6] E. Elhamifar and R. Vidal, "Sparse subspace clustering: Algorithm, theory, and applications," IEEE transactions on pattern analysis and machine intelligence, vol. 35, no. 11, pp. 2765–2781, 2013.

[7] P. Favaro, R. Vidal, and A. Ravichandran, "A closed form solution to robust subspace estimation and clustering," in CVPR 2011. IEEE, 2011, pp. 1801–1807.

[8] M. Soltanolkotabi and E. J. Candes, "A geometric analysis of subspace clustering with outliers," The Annals of Statistics, vol. 40, no. 4, pp. 2195–2238, 2012.

[9] M. Soltanolkotabi, E. Elhamifar, and E. J. Candes, "Robust subspace clustering," The annals of Statistics, vol. 42, no. 2, pp. 669–699, 2014.

[10] C. You, C. Li, D. Robinson, and R. Vidal, "Self-representation based unsupervised exemplar selection in a union of subspaces," IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020.

[11] E. Elhamifar, G. Sapiro, and R. Vidal, "See all by looking at a few: Sparse modeling for finding representative objects," in 2012 IEEE conference on computer vision and pattern recognition. IEEE, 2012, pp. 1600–1607.

[12] C. You, D. Robinson, and R. Vidal, "Scalable sparse subspace clustering by orthogonal matching pursuit," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 3918–3927.

[13] X. Peng, L. Zhang, and Z. Yi, "Scalable sparse subspace clustering," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2013, pp. 430–437.

[14] C. You, C.-G. Li, D. P. Robinson, and R. Vidal, "Oracle based active set algorithm for scalable elastic net subspace clustering," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 3928–3937.

[15] A. Ng, M. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," Advances in neural information processing systems, vol. 14, 2001.

[16] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," The journal of machine learning research, vol. 17, no. 1, pp. 2096–2030, 2016.

[17] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," arXiv preprint arXiv:1412.6980, 2014.

[18] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," Proceedings of the IEEE, vol. 86, no. 11, pp. 2278–2324, 1998.

[19] H. Xiao, K. Rasul, and R. Vollgraf, "Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms," arXiv preprint arXiv:1708.07747, 2017.

[20] M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz, "Invariant risk minimization," arXiv preprint arXiv:1907.02893, 2019.

[21] J. Bruna and S. Mallat, "Invariant scattering convolution networks," IEEE transactions on pattern analysis and machine intelligence, vol. 35, no. 8, pp. 1872–1886, 2013.

[22] "https://www.kaggle.com/c/dogs-vs-cats-redux-kernels-edition," 2013.

[23] Y. Yu, K. H. R. Chan, C. You, C. Song, and Y. Ma, "Learning diverse and discriminative representations via the principle of maximal coding rate reduction," Advances in Neural Information Processing Systems, vol. 33, pp. 9422–9434, 2020.

[24] A. Strehl and J. Ghosh, "Cluster ensembles—a knowledge reuse framework for combining multiple partitions," Journal of machine learning research, vol. 3, no. Dec, pp. 583–617, 2002.

[25] L. Hubert and P. Arabie, "Comparing partitions," Journal of classification, vol. 2, no. 1, pp. 193–218, 1985.

[26] M. N. Akbar, S. F. Ruf, A. Singh, R. Faghihpirayesh, R. Garner, A. Bennett, C. Alba, T. Imbiriba, M. La Rocca, D. Erdogmus et al., "Post traumatic seizure classification with missing data using multimodal machine learning on dmri, eeg, and fmri," medRxiv, 2022.

[27] Q. Wang, W. Rao, S. Sun, L. Xie, E. S. Chng, and H. Li, "Unsupervised domain adaptation via domain adversarial training for speaker recognition," in 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018, pp. 4889–4893.

[28] A. Singh, C. Westlin, H. Eisenbarth, E. A. R. Losin, J. R. Andrews-Hanna, T. D. Wager, A. B. Satpute, L. F. Barrett, D. H. Brooks, and D. Erdogmus, "Variation is the norm: Brain state dynamics evoked by emotional video clips," in 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC). IEEE, 2021, pp. 6003–6007.

[29] L. F. Barrett, "The theory of constructed emotion: an active inference account of interoception and categorization," Social cognitive and affective neuroscience, vol. 12, no. 1, pp. 1–23, 2017.